InFER 2011, Warwick, March 28th- April 1st

Inferring the spatial spread of invasive aliens using epidemiological models

Glenn Marion

Biomathematics & Statistics Scotland - BioSS

www.bioss.ac.uk/

Joint work with:

Stephen Catterall (BioSS)
Alex Cook (National University Singapore),
Adam Butler (BioSS)
Phil Hulme (Lincoln University, New Zealand)
Gavin Gibson (Heriot Watt University)

Funding acknowledgments:

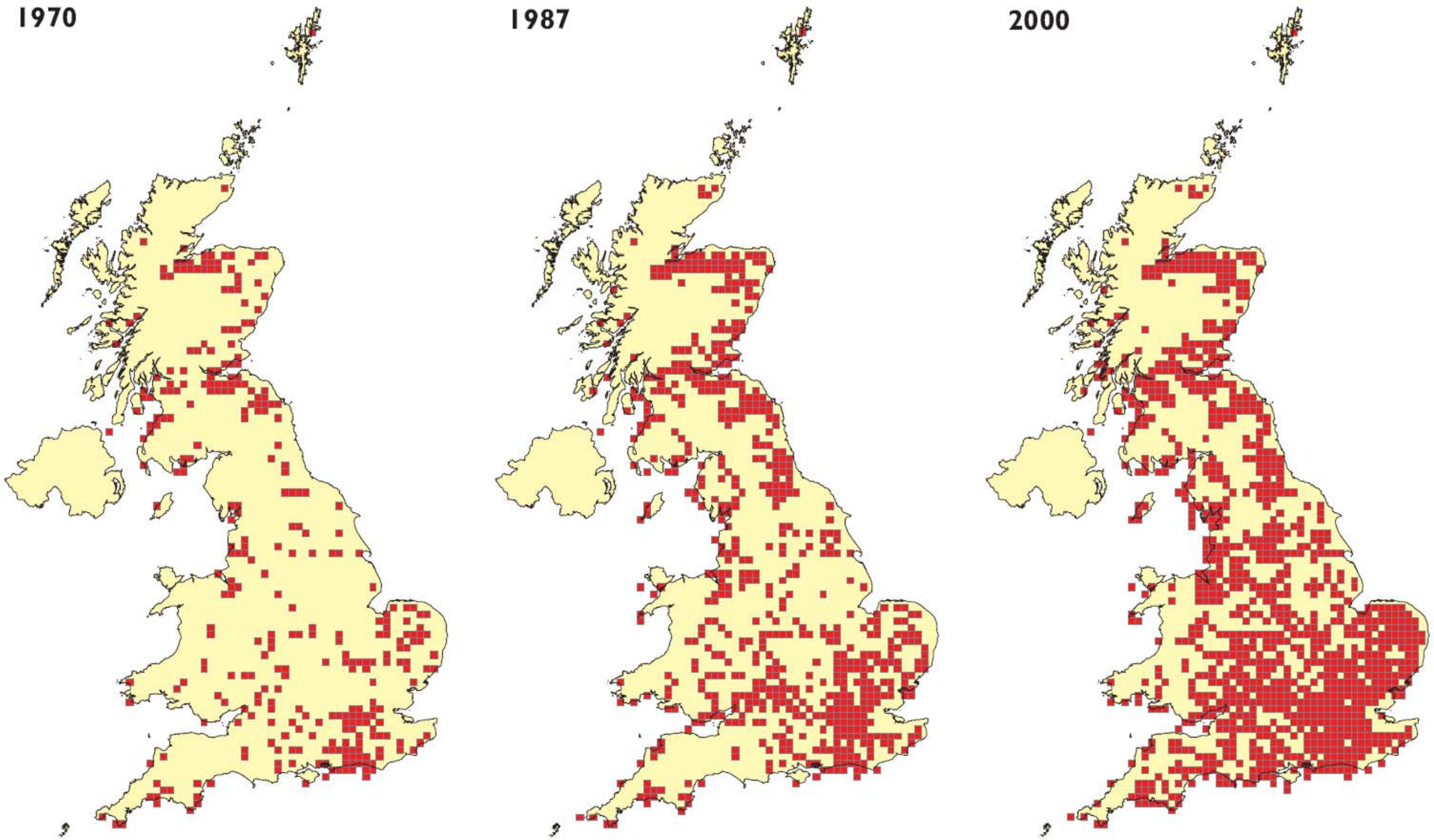
Scottish Government & EU FP6 ALARM project

Spatial spread of invasive aliens

Want to make use of available spatio-temporal data

Species atlas data – provide snapshots at several points in time

BioSS



Observed spread of giant hogweed 1970 – 2000: at hectad scale

But also available for hundreds of other species

Motivation

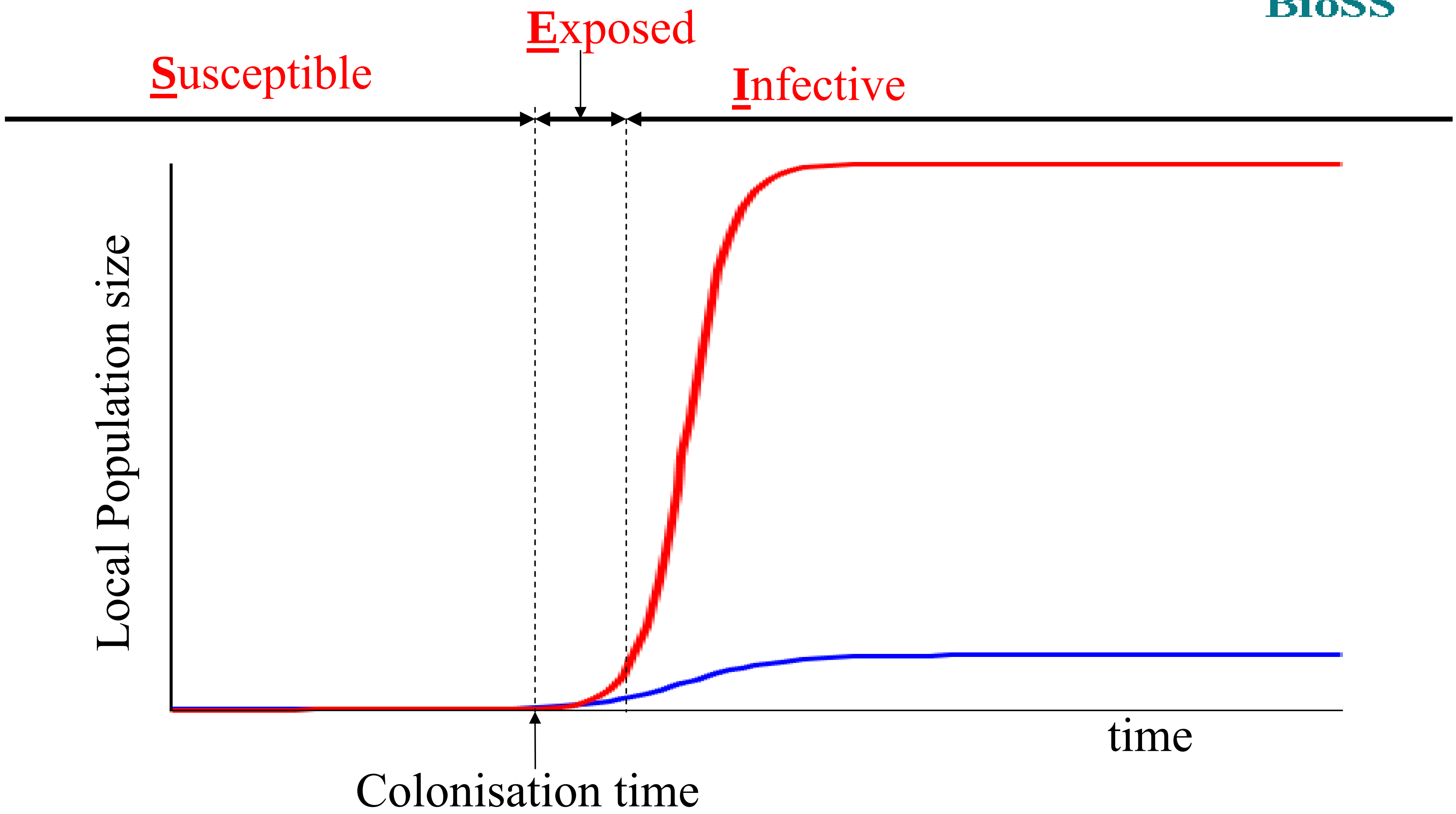
Existing approaches to this type of data:

- Spatial statistical modelling to infer effect of local characteristics (covariates/risk factors) on species distribution
  - tend to ignore temporal dynamics
  - often applied to less easily available data from home ranges
  - relatively quick to apply and uses distribution data
- Detailed mechanistic models
  - time consuming to produce for new species
  - require lots of species specific parameters e.g. dispersal
  - difficult to parameterise using atlas type data
  - more reliable/applicable for some specific questions?

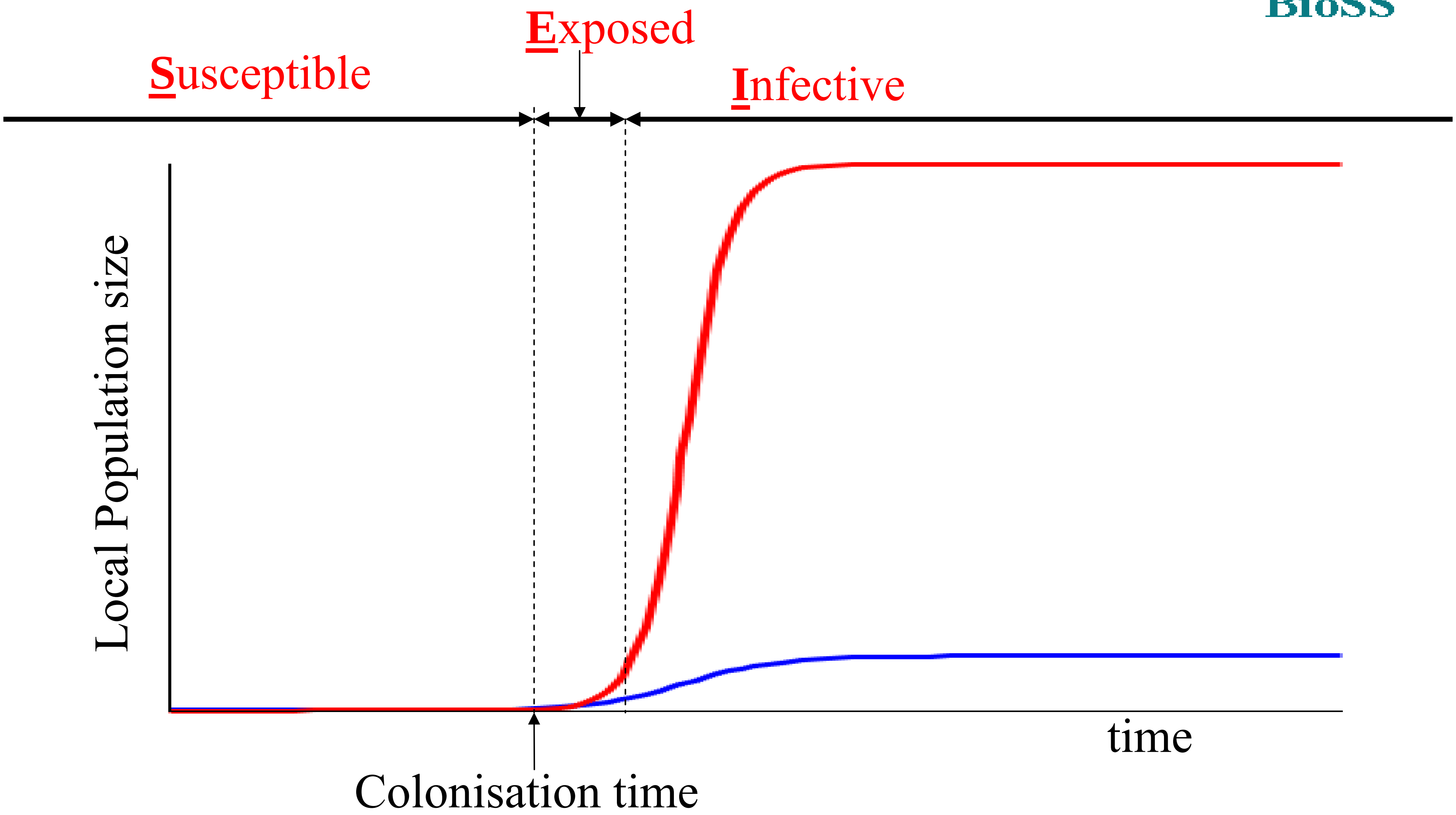
Our aim was to develop

- a generic spatio-temporal (process-based) model
- and routine inference applicable to atlas data and local covariate data

SI or SEI model?

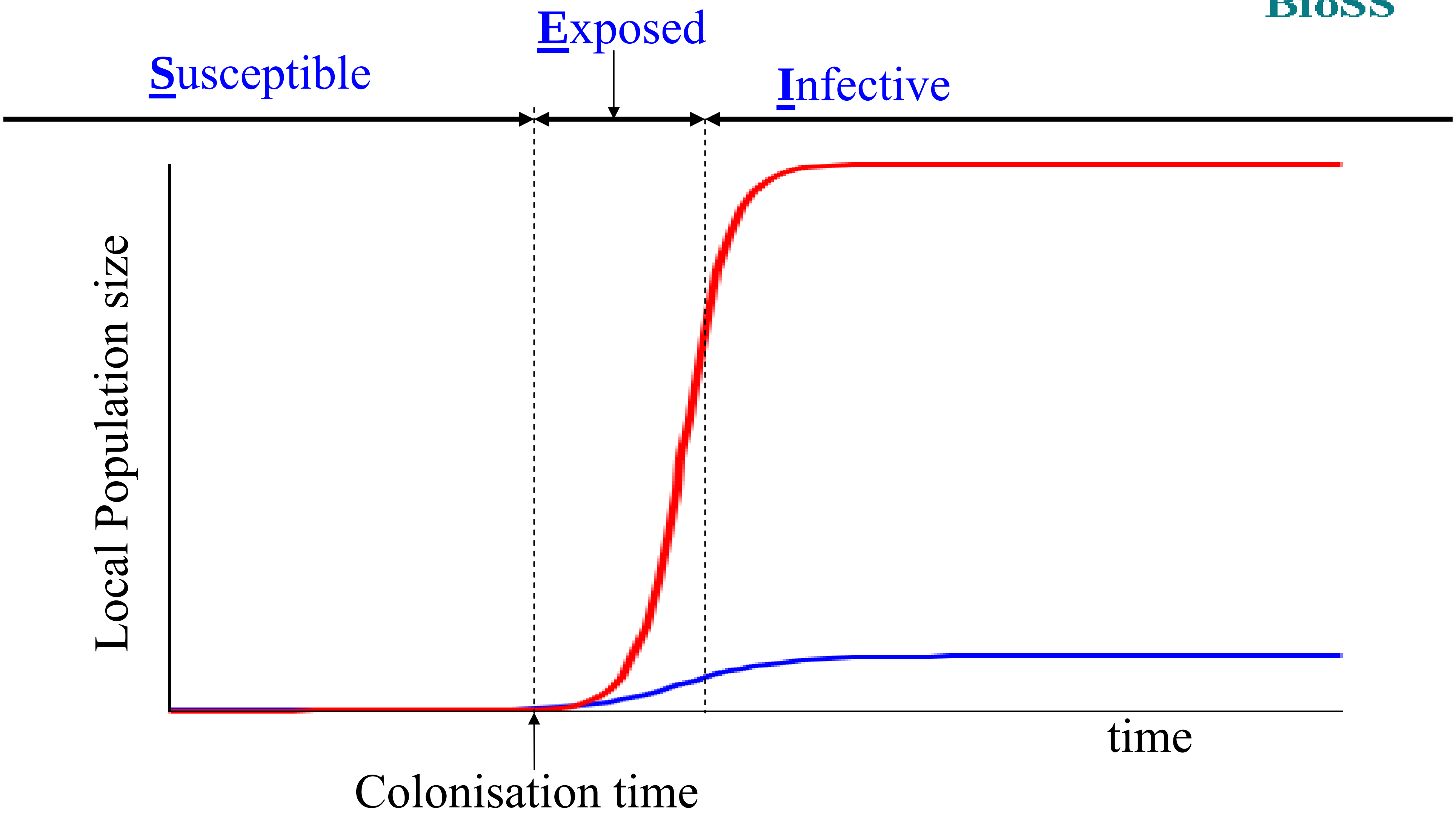


SI or SEI model?

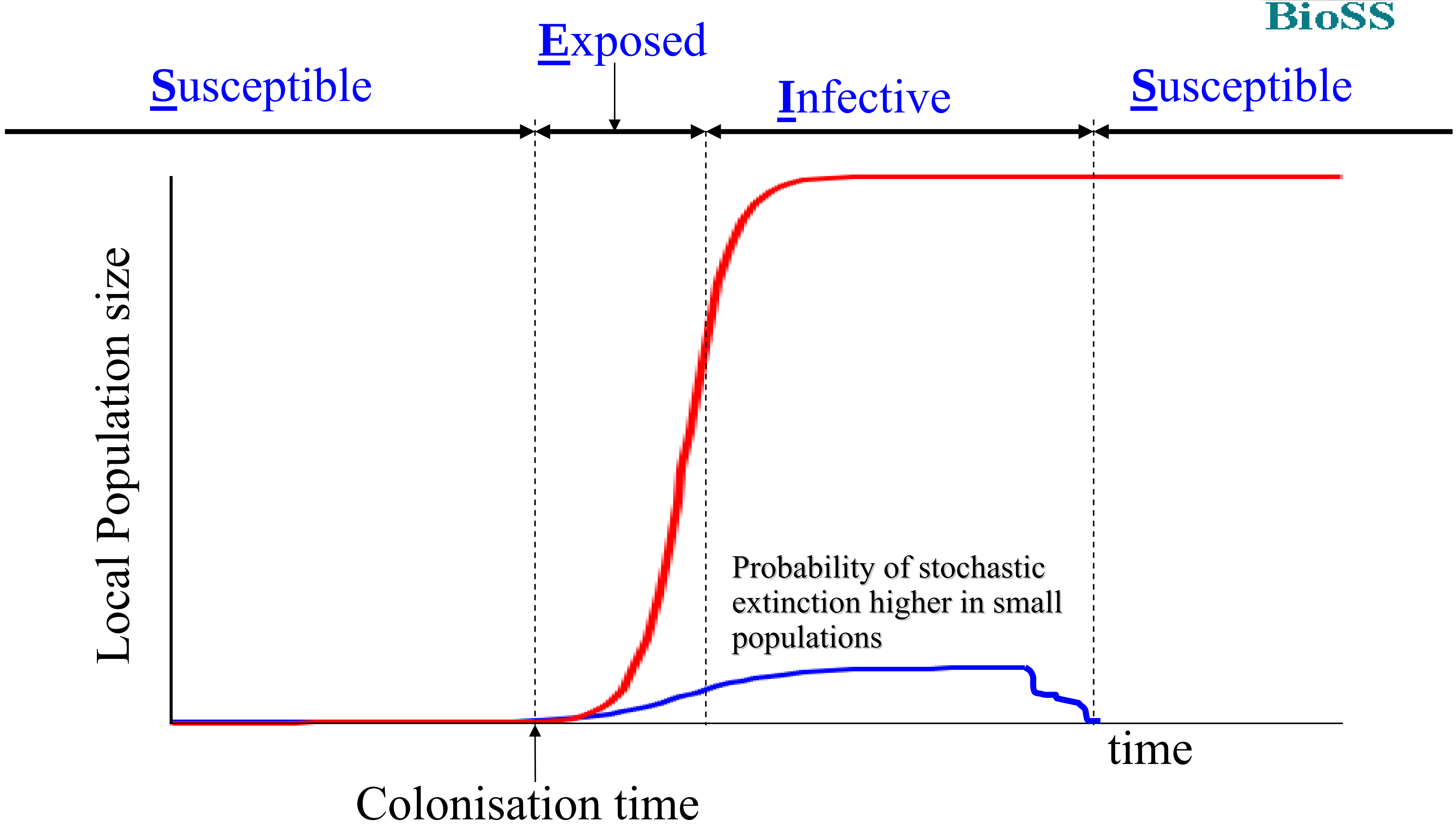


…. could also consider I-low and I-high classes: $SI_L I_H$

SEI model more appropriate



SEI or SEIS model?

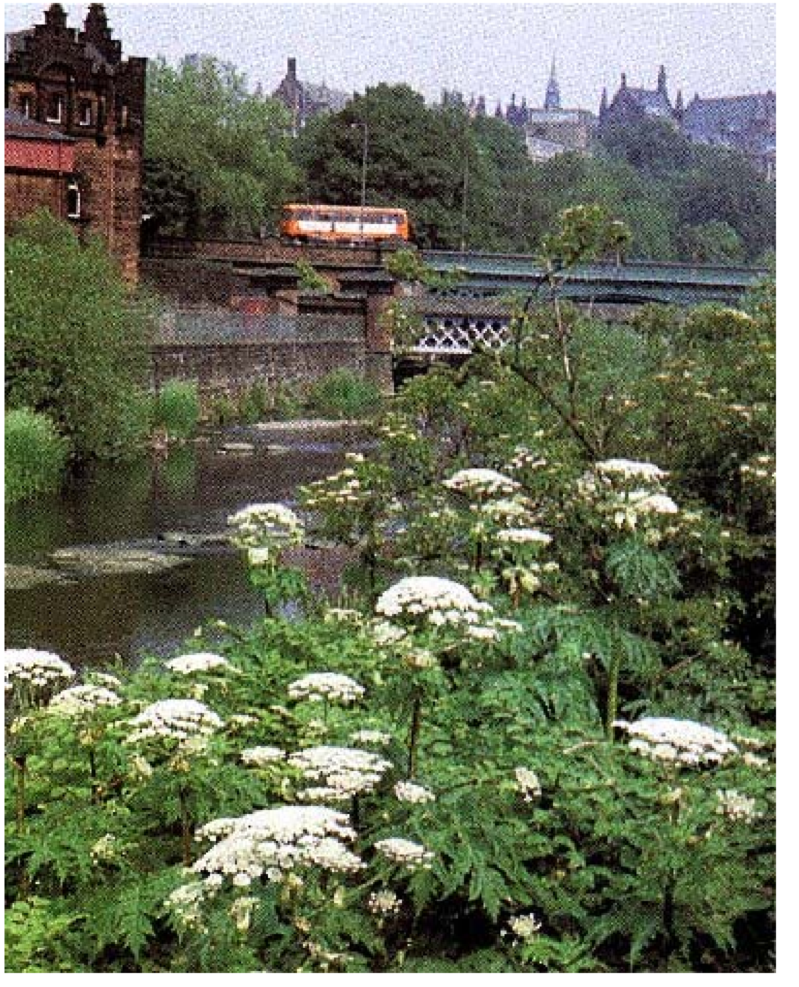


Modelling the spread of giant hogweed

Giant hogweed (*Heracleum Mantegazzianum*)

- Imported 1893, escaped early C20th
- Present in over 35% of 10x10km2 in GB.
- Very difficult to eradicate

- SI model is an appropriate simplification
  - especially at hectad scale

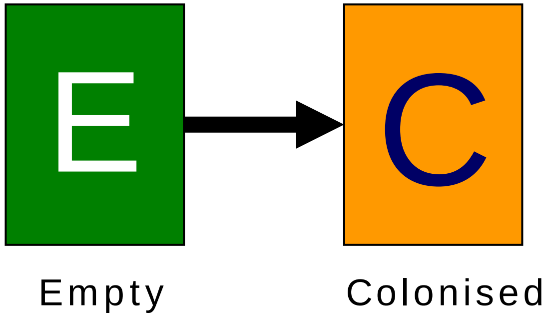


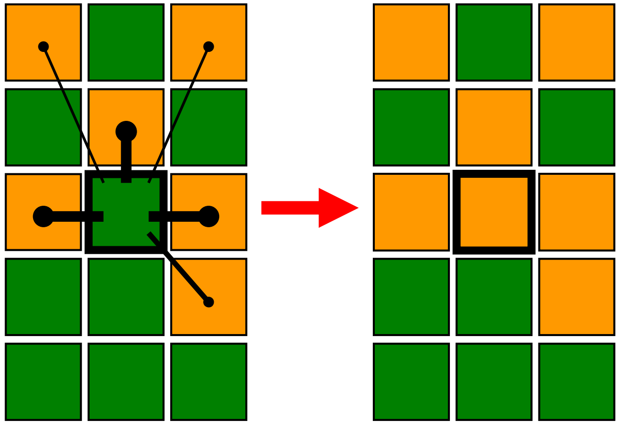
http://www.gla.ac.uk/ibls/DEEB/jd/apfg.html

Modelling assumptions and simplifications

The model: spatial spread

Colonisation rate

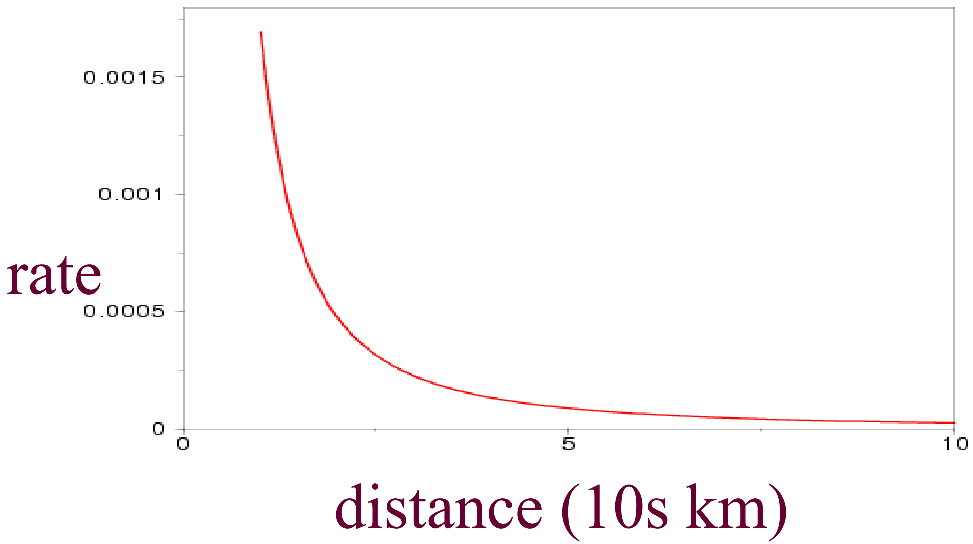




- **Dispersal rate** into uncolonised *i*:

$$\propto \sum_{j} d_{ij}^{-2\lambda}$$

Depends on number of colonised sites
And distances between them



- **Establishment:**

Depends on local habitat and environment in destination patch

The model: establishment probability

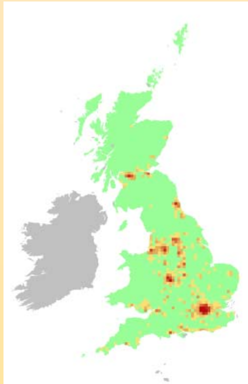
Covariates included as a regression type function of suitability

These are local risk factors for colonisation

Landcover data

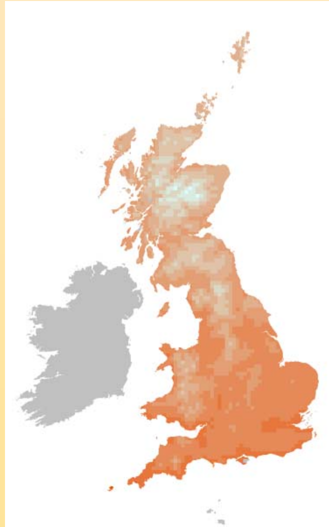
Proportion of land class per hectad of:

Urban, industrial, agricultural, natural, wetland, river, sea, 'unknown'



Climate data

Mean annual temperature over period 1920-2000 per hectad



Wish to carry out inference of dispersal and risk factors based on distribution maps at multiple points in time.

This is possible for this model and also in principle for many natural extensions to it.

Parameter estimation for stochastic processes

Discrete state-space continuous time Markov processes

- state of system: $s(t)$ at time t and q event types $\{e_i : i = 1, ..., q\}$
- type e_i induces a change δs_{e_i} i.e. $s(t) \rightarrow s(t) + \delta s_{e_i}$
- The rate at which event e_i occurs: $r(e_i, s(t); \mathbf{a})$
- The total event rate at time t is $R(s(t); \mathbf{a}) = \Sigma_{i=1}^{q} r(e_i, s(t); \mathbf{a})$.
- The density associated with occurrence of event e_i at $t + \tau$ is:

$$P(s(t+\tau) = s(t) + \delta s_{e_i} \mid s(t)) \;\; = \;\; r(e_i, s(t); \mathbf{a}) e^{-\tau R(s(t);\mathbf{a})}$$

Parameter estimation in stochastic pbms

- The complete set of events $\mathcal{E} = \{(E(k), t_k) : k = 1, ..., n\}$.

- A complete realization of the state-space of the stochastic process reconstructed from $\mathcal{E}$ and $s(t_0)$: $S = \{\mathcal{E}, s(t_0)\}$.

- The *complete likelihood* is:

$$P(\mathcal{E} \mid \mathbf{a}, s(t_0)) \propto \prod_{k=1}^{n} r(E(k), s(t_{k-1}); \mathbf{a}) e^{-(t_k - t_{k-1}) R(s(t_{k-1}); \mathbf{a})}$$

- Follows from model definition

Using data

Suppose have some observations of the state of the system : $\mathcal{D}$

Must define a noise model

$$P(\mathcal{D} \mid S, \mathbf{a}_N)$$

To relate underlying state of the system *S* to the data $\mathcal{D}$

Now can form a **combined likelihood**:

$$P(\mathcal{D} \mid S, \mathbf{a}_N) P(\mathcal{E} \mid \mathbf{a}_P, s(t_0)) = P(\mathcal{D}, \mathcal{E} \mid \mathbf{a})$$

from the complete likelihood and the noise model

A Bayesian approach

- Must specify priors. Typically make independence assumption

$$P(\mathbf{a}) = \prod_{k=1}^{N} P(a_k)$$

- Assume rectangular, or unnormalised flat priors
- or some other standard (or convenient) form e.g. gamma
- Prior distributions can also be specified by previous analyses

Parameter estimation for stochastic processes

Apply Baye's rule to obtain the *posterior distribution*

$$P(\mathbf{a}, \mathcal{E} \mid \mathcal{D}) = \frac{P(\mathcal{D}, \mathcal{E} \mid \mathbf{a}) P(\mathbf{a})}{P(\mathcal{D})}$$

The posterior: P(What we want to know **given** what we do know)

Recall have defined $P(\mathcal{D}, \mathcal{E} \mid \mathbf{a})$ and $P(\mathbf{a})$

But $P(\mathcal{D})$ unknown

However stochastic sampling methods such as Markov chain Monte Carlo (MCMC) allow one to draw samples from the posterior

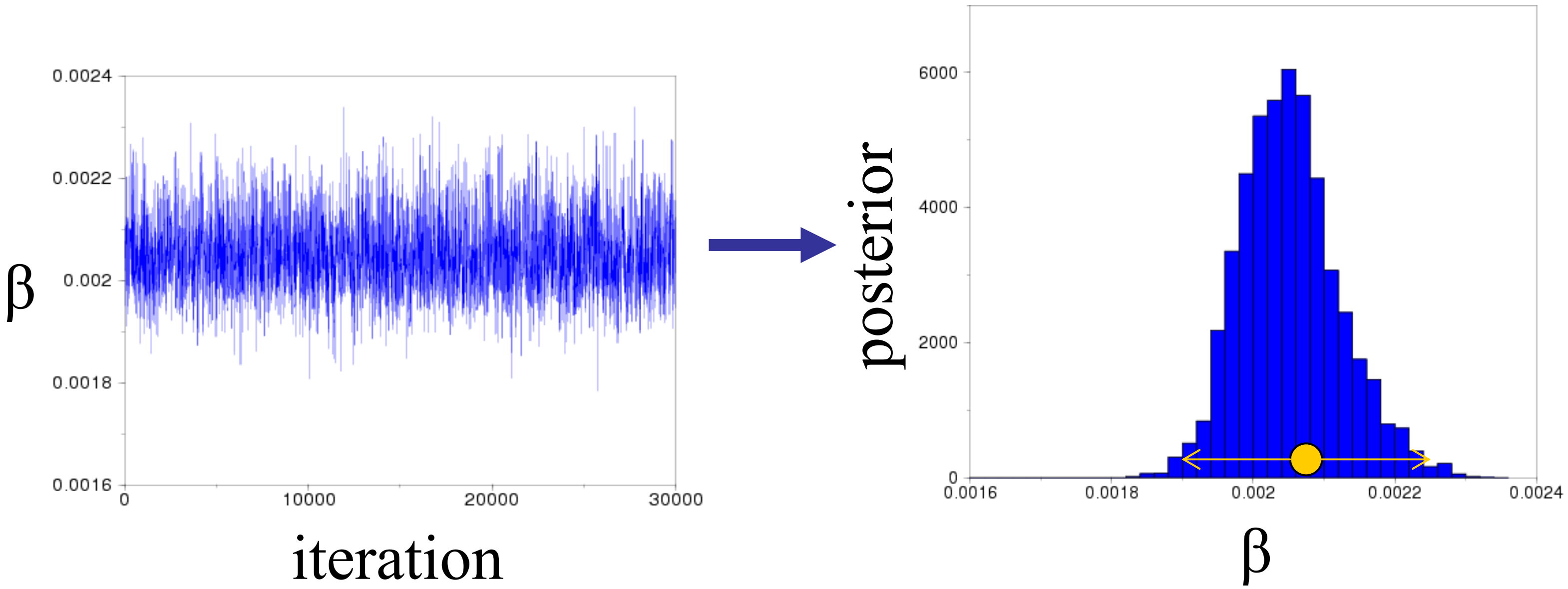
Initialise history consistent with the data & allow chain to converge!

A general inference framework for stochastic process models

Parameter estimation for stochastic processes

Parameter estimation via MCMC computationally intense

- Draw samples from the Posterior distribution
- Samples represent our uncertainty in knowledge - e.g. colonisation times
- Can calculate any statistic of interest from these samples



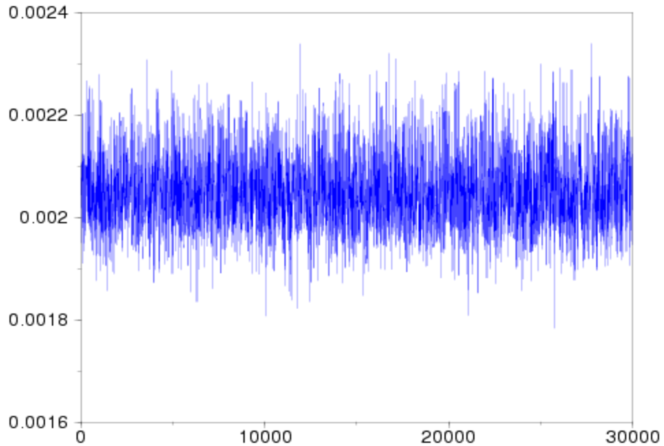
These can include information criterion like DIC for model selection

- trade-off fit to data & model complexity: approximate

Parameter estimation for stochastic processes

Parameter estimation via MCMC computationally intense

- Draw samples from the Posterior distribution
- Samples represent our uncertainty in knowledge - e.g. colonisation times
- Can calculate any statistic of interest from these samples



β

iteration

Also allows outputs in forms more relevant to decision makers

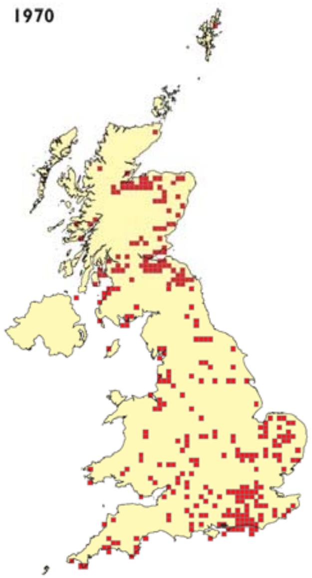
Inference, model and data

Apply this inference framework to:

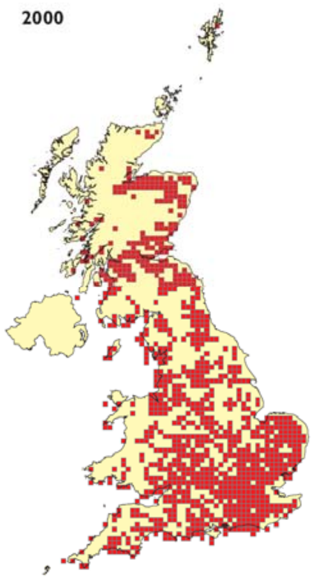
- Colonisation-dispersal model
- Atlas data from 1970 and 2000 – since 1987 data considered unreliable

Infer

- Model parameters: dispersal and colonisation risk
- And unobserved colonisation times



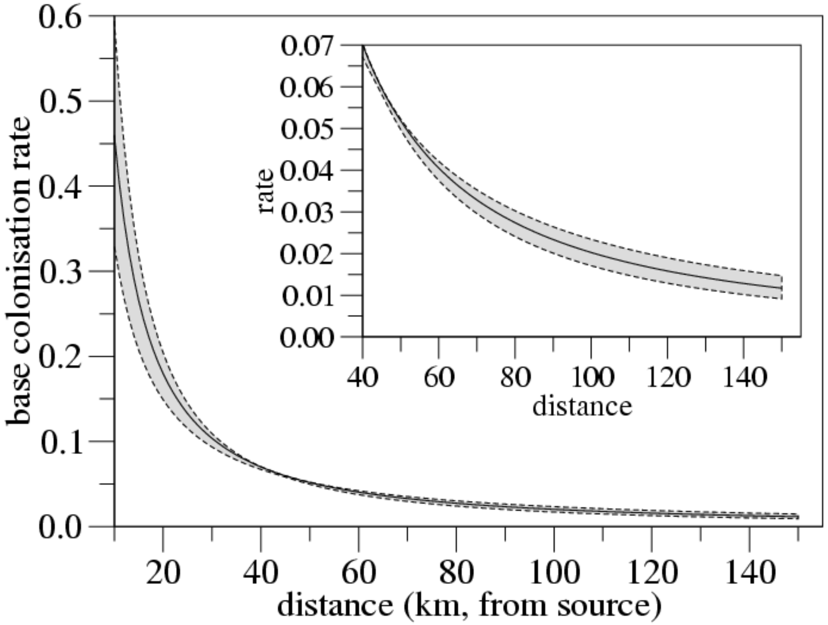
Missing colonisation times?



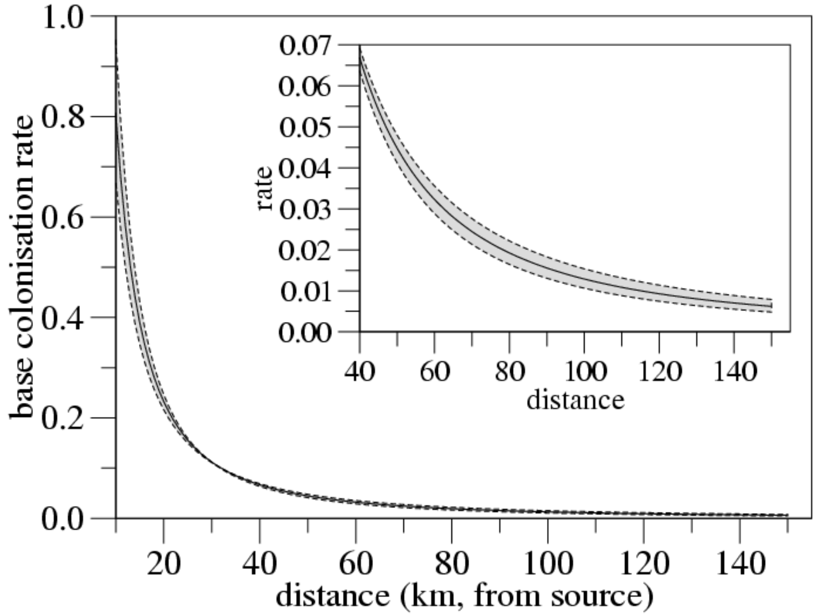
Predictions based on part of process

- Biased estimate of dispersal without covariates
  - longer range dispersal estimated with covariates
- Also biased estimates of covariate effects when ignore dispersal

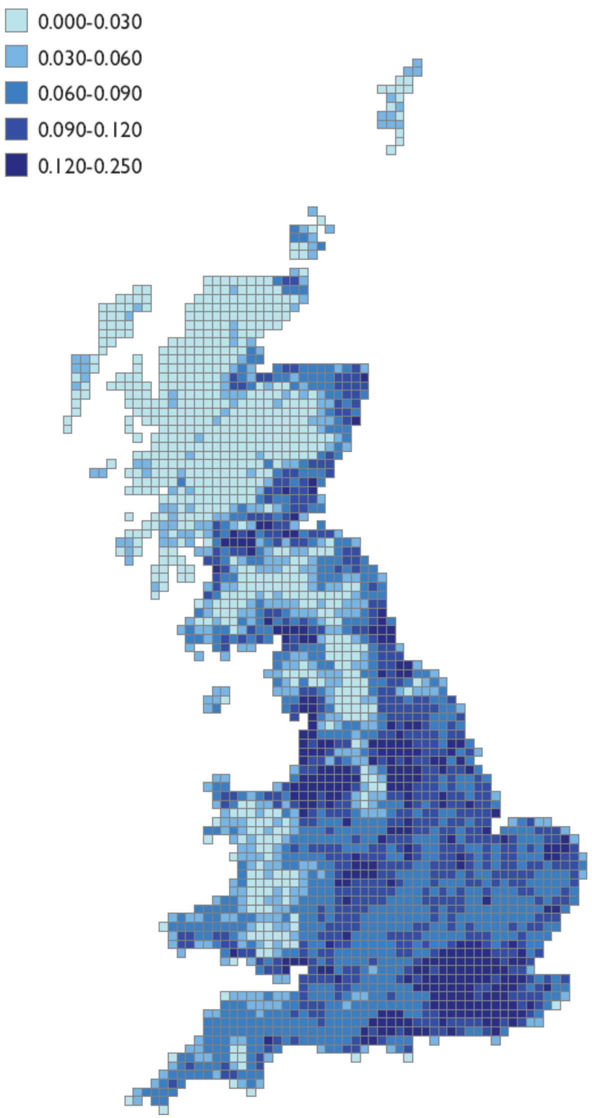
Full model



Null model



Suitability for Giant Hogweed across UK



Inferring unobservable quantities from the atlas data & the model:

Posterior mean colonisability

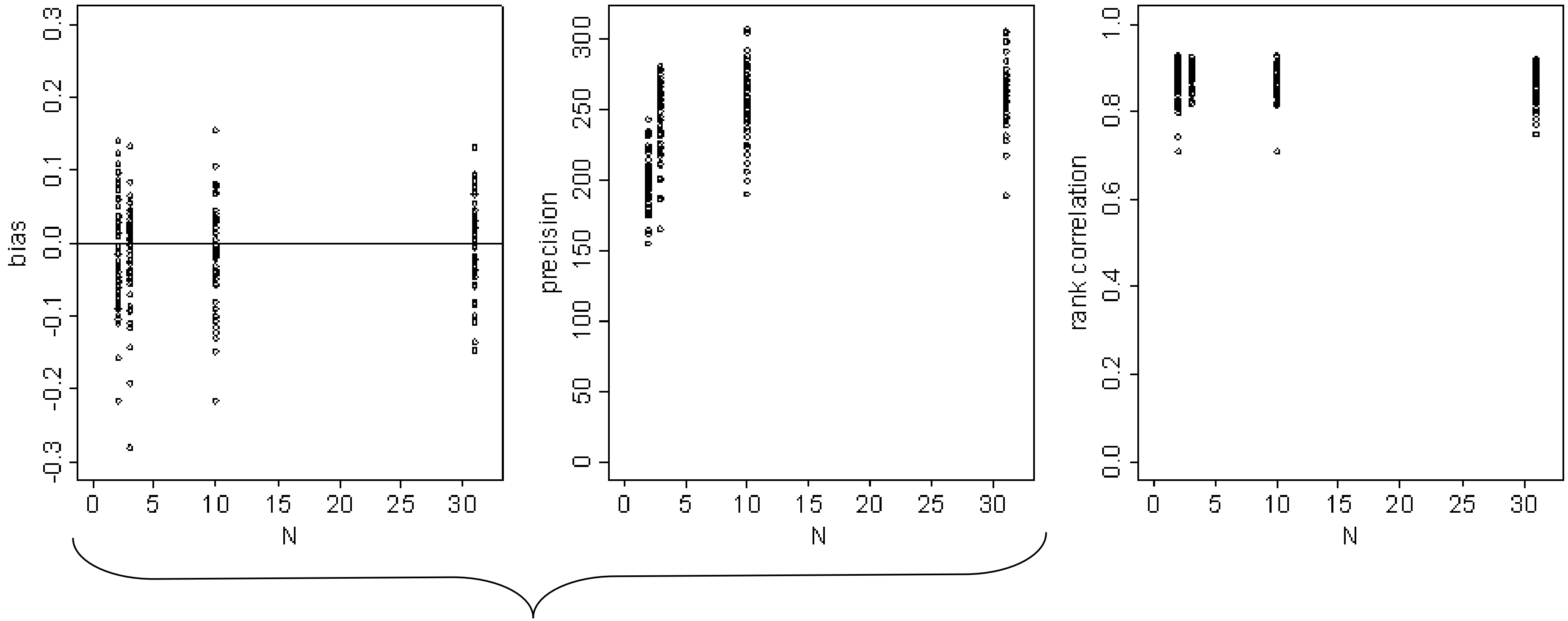
Reliability of inference

Parameter	Estimate and 95% credible interval	Proportion of credible intervals containing original parameter value
λ (dispersal)	1.0 (0.9, 1.2)	93%
α (altitude)	-0.0053 (-0.0071, -0.0034)	82%
τ (temperature)	-0.24 (-0.37, -0.09)	96%
Suitability parameters:		
β_2 (coast)	0.30 (0.10, 0.54)	98%
β_3 (arable)	0.08 (0.05, 0.12)	97%
β_4 (broadleaf forest)	0.72 (0.37, 1.17)	94%
β_5 (urban)	0.41 (0.24, 0.61)	95%
β_6 (conifer forest)	0.04 (0.00, 0.14)	96%
β_7 (improved grassland)	0.16 (0.09, 0.25)	93%
β_8 (open water)	0.15 (0.00, 0.53)	93%
β_9 (semi-natural)	0.09 (0.01, 0.20)	95%
β_{10} (upland)	0.01 (0.00, 0.05)	98%

Reliability of inference assessed using simulated data generated from posterior mean estimates of parameter values for giant hogweed.

Increasing the number of snapshots

The number of observation time points has little impact on quality of inference

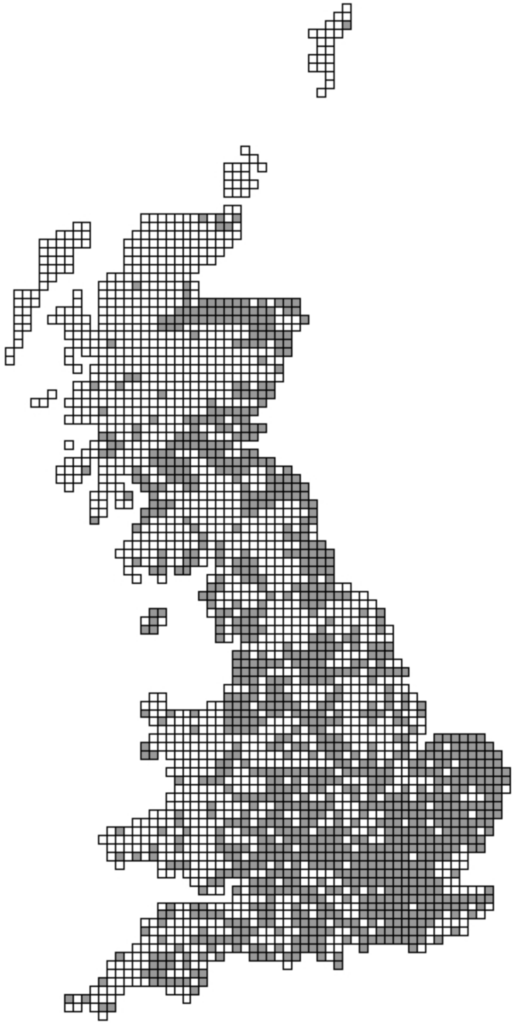


Dispersal parameter

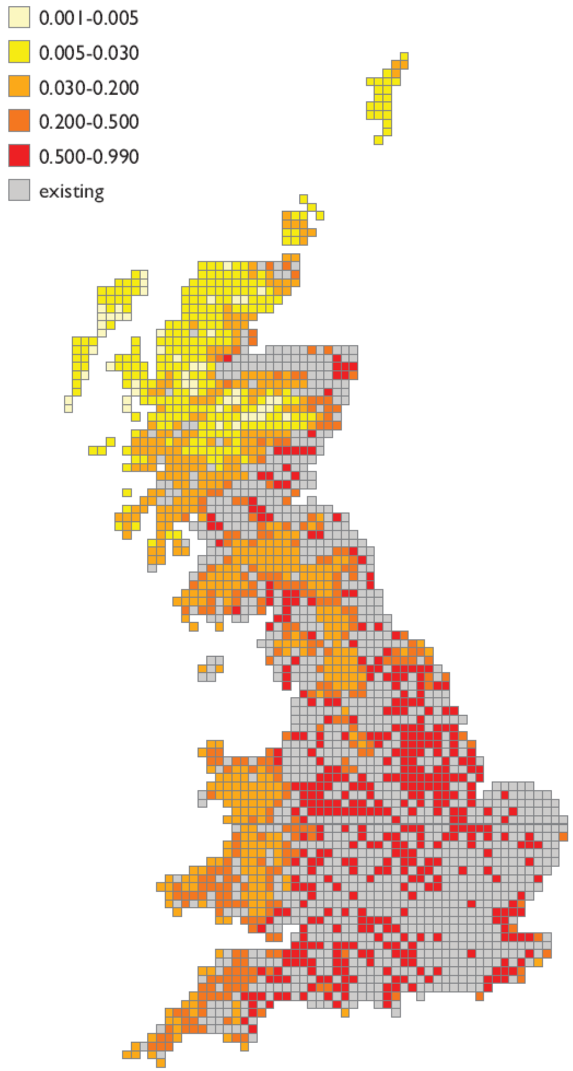
Risk factors

Inference and the spread of invasive aliens

Predictions: uncertainty & variability



Variation between individual runs reflects uncertainty



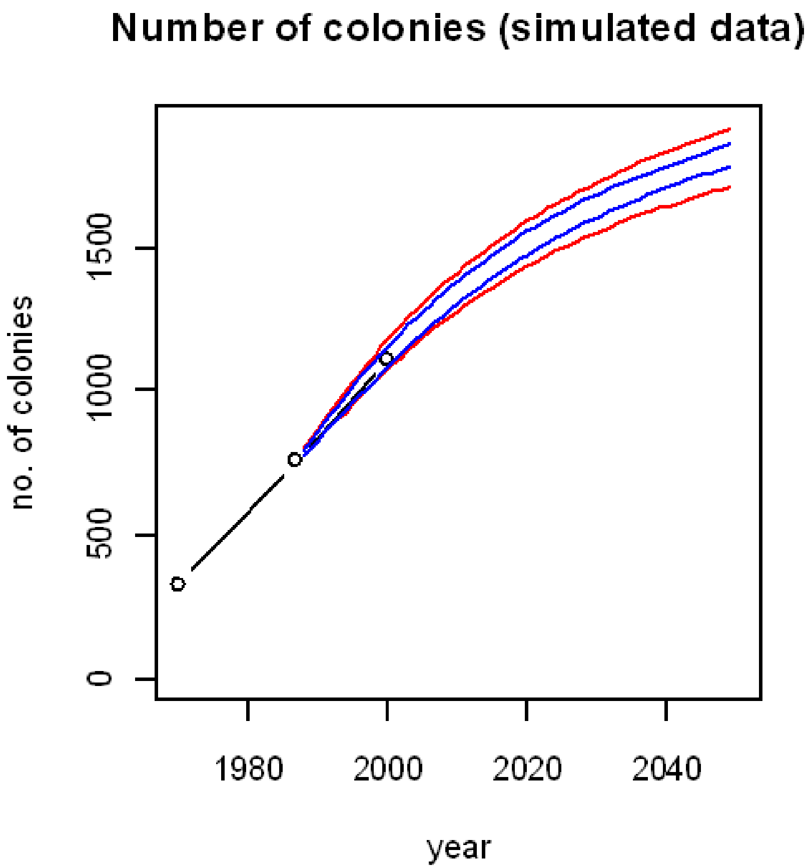
Colonisation probability in 2050 across all runs – explore different scenarios

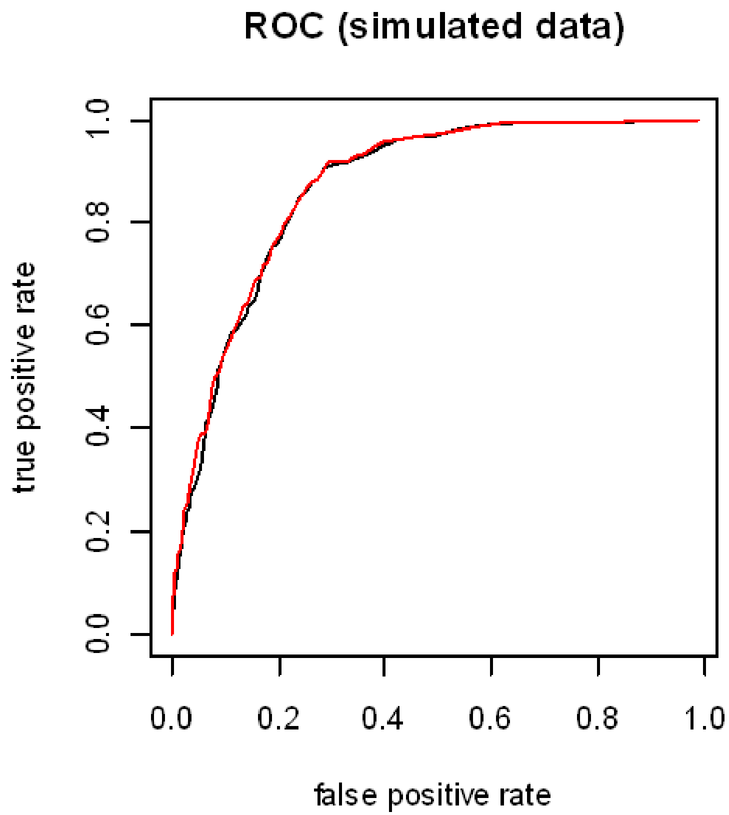
Example: modelling the spread of invasive aliens

Testing prediction

Based on simulated data

Little loss of predictive power due to uncertainty in parameter estimates



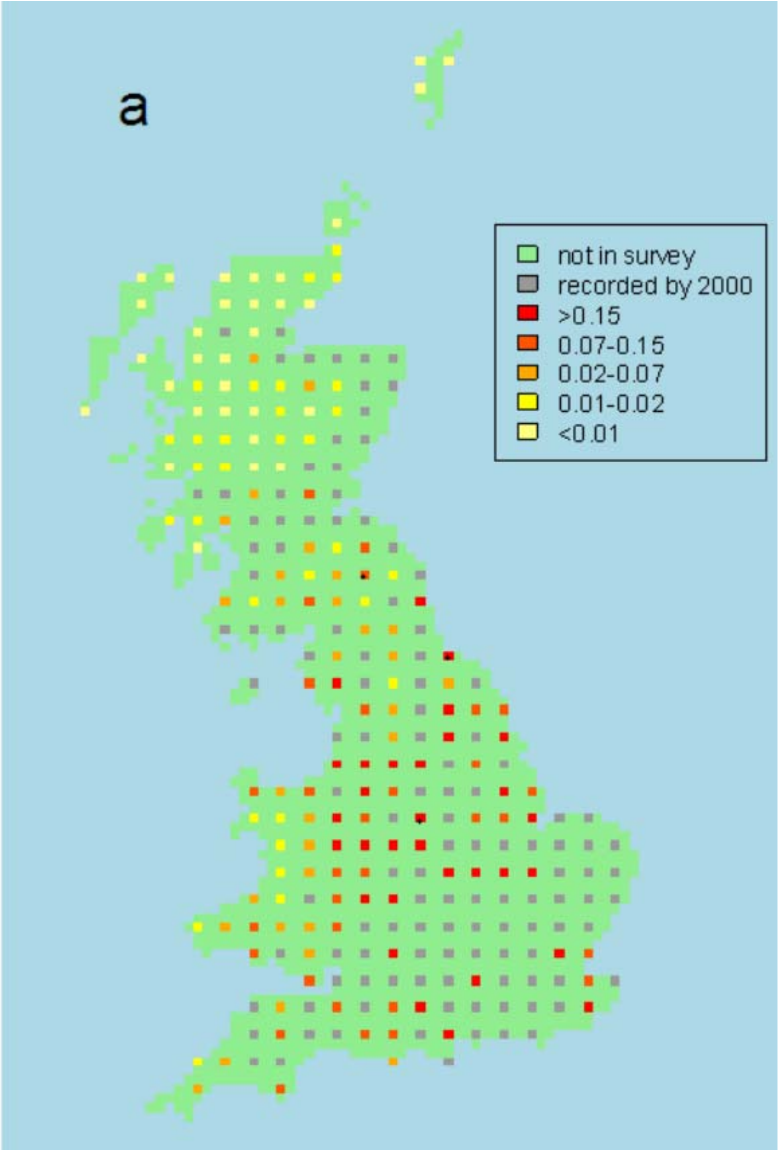


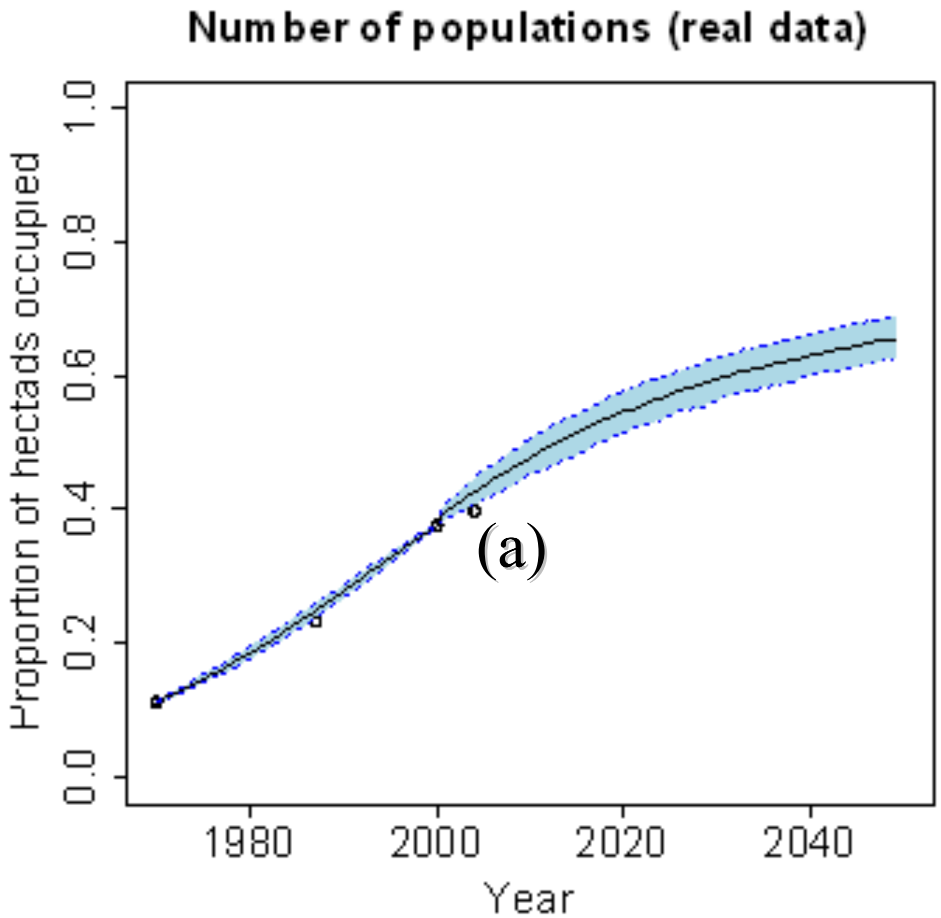
Based on knowledge of parameters values (blue) and inferred parameter values (red)

Testing prediction

… based on inference from maps at 1970 and 2000

a) Limited survey of 1 in 9 hectads in 2004

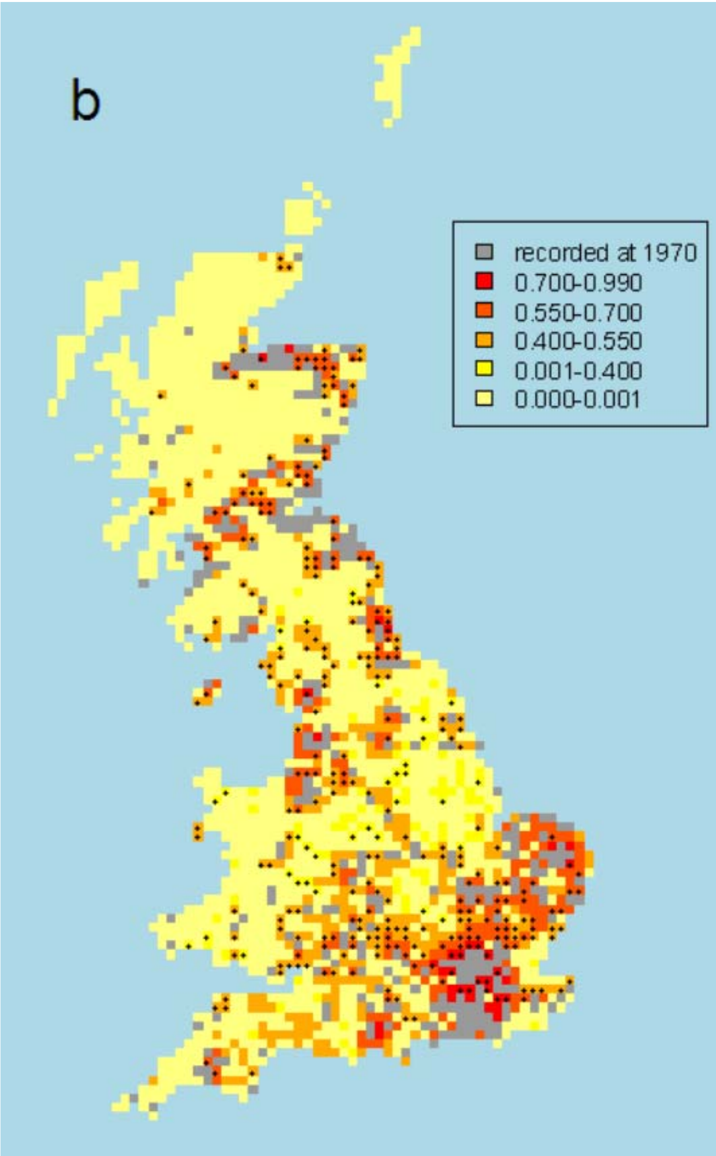


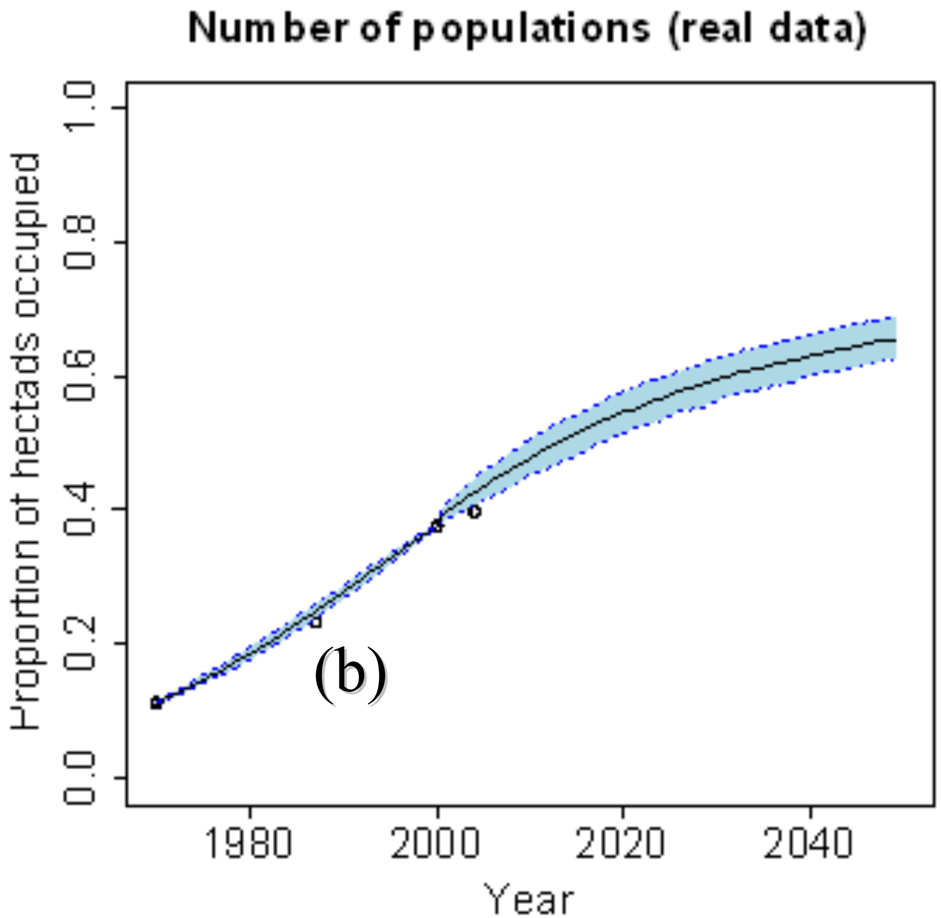


Testing prediction

… based on inference from maps at 1970 and 2000

b) 1987 atlas which is known to have underreporting bias.



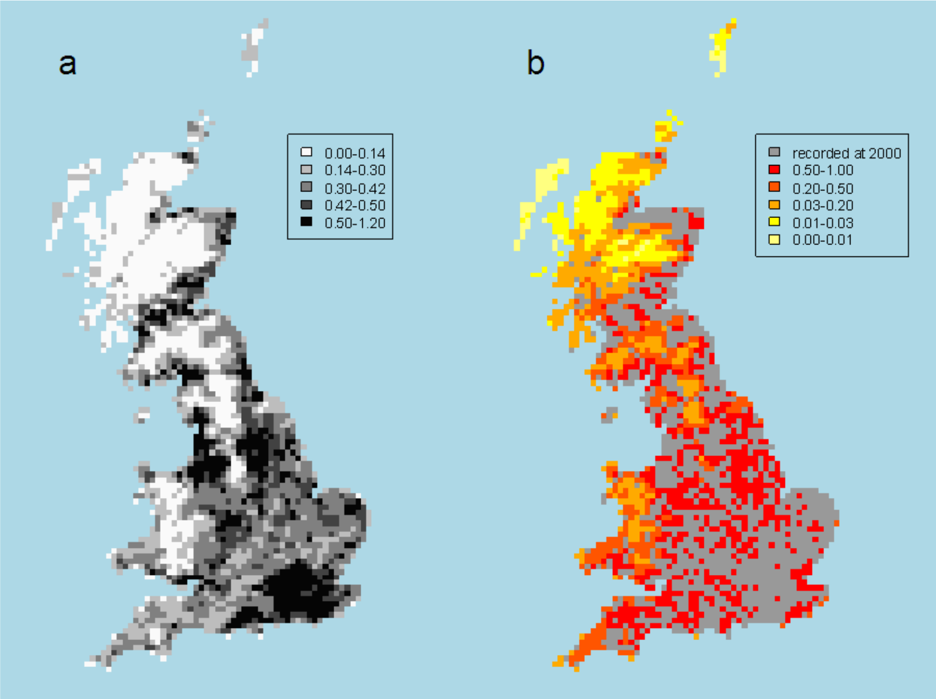


Example: modelling the spread of invasive aliens

Future applications

Could use this framework to:

- Predict future spread under different climate scenarios
- Target high risk areas for control
- Assess the impact of different interventions
- Infer characteristics of a wide range of invasive species with a view to identifying invasive traits.



Example: modelling the spread of invasive aliens

Methodological developments

Extend this framework to

- Account for inhomogeneous detection rates
- Model range shifting species (SIS model)
- Model abundance and local population dynamics
- Account for temporal variation in covariates (risk factors) e.g. landuse change

Parameter estimation in stochastic pbms

- The complete set of events $\mathcal{E} = \{(E(k), t_k) : k = 1, ..., n\}$.
- A complete realization of the state-space of the stochastic process reconstructed from $\mathcal{E}$ and $s(t_0)$: $S = \{\mathcal{E}, s(t_0)\}$.
- The *complete likelihood* is:

$$P(\mathcal{E} \mid \mathbf{a}, s(t_0)) = \prod_{k=1}^{n} r(E(k), s(t_{k-1}); \mathbf{a}) e^{-(t_k - t_{k-1}) R(s(t_{k-1}); \mathbf{a})}$$

- Follows from model definition